Use of Taxi-Trip Data in Analysis of Demand Patterns for Detection and Explanation of Anomalies

Markou, Ioulia; Rodrigues, Filipe; Pereira, Francisco Camara

Published in:
Transportation Research Record

Link to article, DOI:
10.3141/2643-15

Publication date:
2017

Document Version
Peer reviewed version

USE OF TAXI-TRIP DATA IN ANALYSIS OF DEMAND PATTERNS FOR DETECTION AND EXPLANATION OF ANOMALIES

Ioulia Markou
PhD Student, DTU Management Engineering, Transport DTU
Bygningstorvet 116B, DK-2800 Kgs. Lyngby
Tel: +45 45251515; Email: markou@dtu.dk

Filipe Rodrigues
Postdoc, DTU Management Engineering, Transport DTU
Bygningstorvet 116B, DK-2800 Kgs. Lyngby
Tel: +45 45256530; Email: rodr@dtu.dk

Francisco C. Pereira
Full Professor, DTU Management Engineering, Transport DTU
Bygningstorvet 116B, DK-2800 Kgs. Lyngby
Tel: +45 45251496; Email: camara@dtu.dk

Word count: 5,497 words text + 8 tables/figures x 250 words (each) = 7,497 words

March 15st, 2017

ABSTRACT

Due to environmental and economic stress, strong investment exists now towards adaptive transport systems that can efficiently utilize capacity, minimizing costs and environmental impacts. The common vision is a system that dynamically changes itself (the supply) to anticipate traveler needs (the demand). In some occasions, unexpected and unwanted demand patterns are noticed in the traffic network that lead to system failures and cost implications. Significantly low speeds or excessively low flows at an unforeseeable time are only some of the phenomena that are often noticed and need to be explained for transport system's better future response.

The objective of this research is the formulation of a proper methodology that identifies anomalies on traffic networks and correlates them with special events using internet data. Our main subject of interest is the investigation of why traffic congestion is happening as well as why there are demand fluctuations in days were there are no apparent reasons for the occurrence of such phenomena. We evaluated our system using Google's NYC taxi trips public dataset. We defined initially the "normality" baseline and thereunder we studied individual days' demand patterns for outliers' detection. Our approach enabled us to detect demand fluctuations, analyze and correlate them with disruptive events scenarios like extreme weather conditions, public holidays, religious festivities and parades. Using kernel density analysis, the affected areas as well as the significance of the observed differences compared to the average day are depicted.

Keywords: Anomaly, Events, Kernel Density, Demand, Taxi Trips

INTRODUCTION

Transport systems function generally well. In some occasions though, unexpected and unwanted performance patterns are noticed that lead to system failures and cost implications. Significantly low speeds or excessively low flows at an unusual time are only some of the phenomena that may confuse a driver or transport authorities, since they are totally unexpected and frequently there is no obvious explanation for them. The term "anomalies" refers to those non-conforming patterns which appear into a well-defined notion of normal behavior. In the literature, similar phenomena can be described as outliers, exceptions or discordant observations. The common feature of all these terminologies is that they represent critical information in a wide variety of application domains, which is particularly useful for momentous events identification and crisis management.

Anomaly detection is extensively used in a wide variety of applications. It is a crucial task in many safety-critical environments, such as fraud detection for credit cards, insurance or health care, intrusion detection, activity monitoring through mobile phones, etc. Anomalies could occur due to changes in the behavior of systems, human errors, natural deviations in populations or fraudulent behavior. The application area typically defines the anomaly detection system and the methodology tools that will effectively identify and collect the necessary information for events assessment.

Transportation networks present several anomalous situations of particular interest and merit. Phenomena of different scale and influence have attracted the interest of several researchers who try to monitor and explore their specifications. Accidents, protests, celebrations, concerts, sport events define crowds, disruptions, road closures, etc., which subsequently cost time, money and urban pollution. Therefore, several methodologies have been developed for the detection and analysis of location, time and the purpose of them, in order to provide improved guidance to the users of the traffic system and reduce the impact of the associated problems. The requirement of understanding why people are travelling on the other hand, is often abundant in natural language form, and has been largely neglected.

Sensors can detect and accurately measure traffic congestion, flow models can represent how it should evolve on a specific network area and time window, but they cannot find a parade organized nearby. Figure 1 shows how New York City Halloween Parade affects taxi demand at the surrounding areas. Regions which have been marked with red shades indicate areas with increased demand for taxi trips. Even when this context is captured manually, limited options exist to correlate the two phenomena: one described with rich semantic information, the other with traffic data. In general, travel choices are strongly context-dependent, but context has been considered difficult, if at all possible, to capture, let alone be included in behavior models. Travel surveys usually don't go further than multiple-choice questions (e.g. travel purpose, accompanying travelers, perceived comfort, weather) because of user burden and cognitive limitations (e.g. memory, context-awareness), which is constraining a whole field to a myopic view of travel behavior.

In previous works in collaboration with the Land and Transport Authority of Singapore, we showed that even simple event information (e.g. event category) collected from an online event directory, can be used to improve public transport arrival predictions *(1)*. However, due to the complexity of the exploration of the open Web (e.g. using Google search), the use of internet data in transportation is currently limited to manually defined sources and highly fine-tuned processes. As mentioned in *(2),* the grand challenge is to break this "barrier" and start jointly considering all kinds of contextual information by broadening the search space to the entire Web, instead of just focusing on a single type of contextual resource such as incident feeds or a manually build list of

event websites.

The main contributions of this paper are the formulation of a proper methodology that identifies traffic anomalies on traffic networks and correlates them with special events using internet data. Our main subject of interest is the exploration of why traffic congestion is happening as well as why there are demand fluctuations in days were there are no apparent reasons for the occurrence of such phenomena. The present study is not yet about automatically searching from the web for a random event, instead it's about getting the first building blocks for this endeavor: automatically detect time, location, and magnitude of such events for real-time explanation of traffic congestions or road closures which are highly correlated with them. Utilizing the developed methodology, we are gradually led to the real-time surveillance of the state of the transportation network during non-recurrent scenarios, such as the events described in the research, and provide real-time information and guidance to travelers and transportation administrators.

The remainder of this paper is structured as follows. A literature review on traffic anomalies and the developed techniques of identification is presented next, leading to the description of the adopted methodology. The setup of the experiments, using public datasets is presented next. A concluding section discusses the main findings and provides directions for further research.

FIGURE 1 Taxi Demand for Halloween Parade NYC 2015

LITERATURE REVIEW

During the last decades, research and development in intelligent transportation systems (ITS) have given mature tools for monitoring, estimation and control of traffic networks *(3)*. They are largely supported by the ubiquity of pervasive technologies, such as radio-frequency identification, GPS, Wi-Fi, NFC and mobile phone communications. All these tools enable researchers to understand the dynamics of a city and improve traffic management and decision making processes.

There is no well-defined threshold above which we identify anomalies on transport systems. They are caused by accidents, sport events, parades, demonstrations, extreme weather conditions etc. The intuition is that anomalies should happen whenever the supply (e.g. buses,

trains, network capacity) is misaligned with the demand (e.g. travelers) in ways that are not common for that location and time.

Anomaly detection could be implemented using either macroscopic or microscopic traffic variables. Variables from the first category, such as flow and occupancy, have been extensively examined for traffic incident detection *(4)*. Speed variations incorporate useful information *(5)* resulting in the sensitivity increase to traffic patterns deviation. The second category of variables describes individual vehicles behaviors which are also valuable for certain research areas. Lane changing fractions, relative speeds and inter-vehicle spacing are some of the parameters that have been studied *(6-8)*. Simulation results showed promising results on transient anomalies and incident detection with low false alarms rates *(8)*. However, the scale of the analysis is not the only parameter that is taken into consideration. Some researchers choose to study the problem from a different perspective, namely to focus on the supply components of a traffic network, such as traffic flows, densities and routing behavior *(9-11)*. Traffic dynamics through segment densities were tried to be understood and used for prediction *(12)*. Topological variation in traffic flow between points and the visualization of the affected road segments of the anomaly has also been studied *(9)*.

Meanwhile, anomalies occurrence is also connected with demand components, and more specifically with non-habitual overcrowding scenarios, such as public special events (sports games, concerts, parades, sales, demonstrations, and festivals) that directly affect them. Transport systems are generally designed with reasonable spare capacity in order to cope with those demand fluctuations. However, in several cases high waiting times are observed due to a congested traffic network that is no longer able to serve increasing transportation and mobility needs. Relevant scenarios have been extensively studied *(13-15)*. The start time as well as the duration of an event are two of the most important parameters that will define how the demand around a certain area would be affected. As a result, several studies are oriented to new data sources that can provide transportation systems and models with that information. By better understanding why these crowds occur, transportation models could be improved and present better planning and prediction results *(16)*.

The exploitation of the information deducted from the Internet attracts great interest *(17-22)*. Social Media (i.e. Facebook, Twitter, Google+ and Flickr) are rich in local context information generated by large online crowds. Information about public special events from social networks and other platforms that have dynamic context content (e.g. news feeds), can help discerning explanations about real-world phenomena.

Generally, anomalies can be seen as a group of observations lying (considerably) outside a region of likely expected values, given the "normal" behavior of a system. Their detection and identification involves continuous estimation of models of normal system behaviors at specific areas or time of interest. This process requires finding the best methodology for the type of data available as well as for the computational capacity of the system. Consequently, researchers developed several anomaly detection techniques to meet these diverse needs. Classification based techniques define abnormal sets of values by inspecting if they exceed a certain threshold. For example, Zhang *(23)* computed shortest paths and compared the recorded distances with them, and Castro et al. *(12)* characterized a road segment as congested when the observed density was above a certain value. Clustering based techniques are based on the assumption that normal data instances belong to a cluster. Bu et al. *(24)* monitored distance based anomalies using data structures and algorithms employing local clustering and Candia et al. *(25)* showed that anomalous events give rise to spatially extended patterns. Statistical based techniques give low probability to anomalous situations. Parametric distributions, histograms, regression models etc., are included in

this category. Finally, information theoretic techniques analyze the information content of a data set and try to minimize its subset size, by simultaneously aiming to the minimum possible information loss. *(26).* Examples of popular algorithms include Latent Dirichlet Allocation, LDA *(27),* which re-represents each document as a linear combination of latent bag-of-words (BoW) vectors, or topics, that can be seen as the building blocks of all documents in the collection. These have had tremendous success, including in transportation applications (e.g. decomposition of an incident record into its probable constituents *(16);* text analysis for special events *(28)*, and urban planning *(29)*.

Above studies show that there is great potential to use internet data for information about planned events and their popularity. However, none of the previous studies explores automatic ways of time, location and magnitude detection of events within an area where a serious traffic congestion or road closure occurs. Additionally, taxi demand distribution has not been used yet for special events identification and explanation.

METHODOLOGY

Definition of normality

The range of the affected area varies during an event and descriptive boundaries cannot be provided easily. A multitude of sources provide high resolution information from alternative channels, such as cellphone/communication data, social platforms information and media coverage combined with the more traditional transport sensor systems (loop detectors, traffic radars, cameras, Wi-Fi and Bluetooth sensors). Depending on the scale of influence, the best processing tools as well as the necessary data precision could be defined for the explanation of abnormal observations

Figure 2 shows an overview of the proposed methodology. It starts with the definition of a "normality" baseline according to historical mobility data, which, in the case of our experiments, corresponds to NYC GPS taxi trips. The first stage of analysis will be conducted by exploring the available data in an extensive time interval, such as a year, or several months, so that we can primarily understand the dynamics of our studied area. The knowledge of which areas present high daily demand, or which days present fewer trips in total, due to low demand from the city residents (like Sundays compared to the remaining days of the week) will restrict incorrect assessments of significantly high or low values of our anomaly indicator parameters.

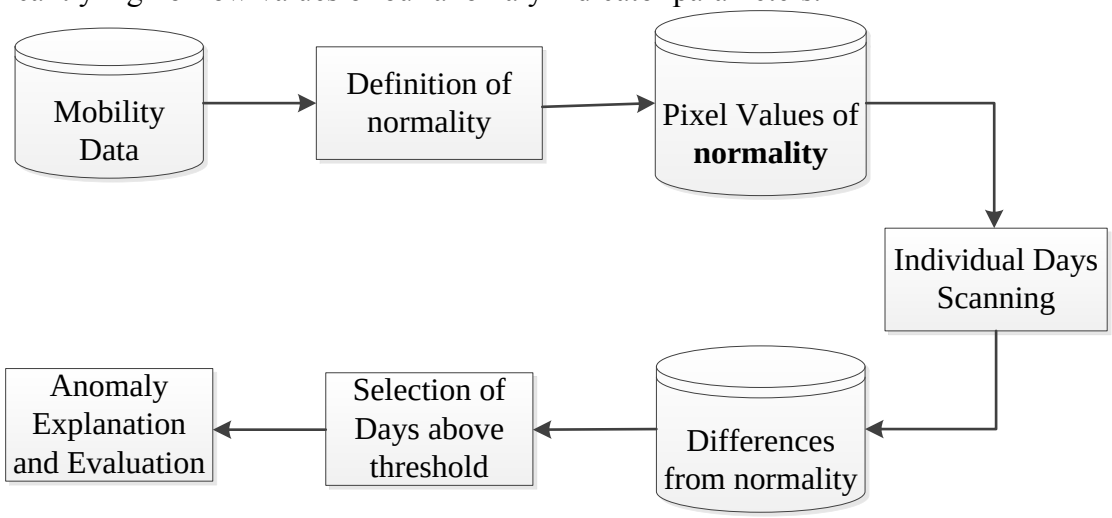


FIGURE 2 Developed Methodology

Kernel Density Estimation

Special events, protests, politicians visit etc. attract a large number of people in a certain area for a short or a long time period. A theatre, a stadium or an exhibition center emerge as point of interest (POI) for a great number of citizens, therefore they influence accordingly the overall heat map of city's daily trips. For the detection of unusual motions or interactions it is necessary to build representations of the selected area of interest that demonstrate regions where the number of trips has significantly changed. A quite general nonparametric technique that estimates the underlying density, thereby avoiding having to store the complete data, is kernel density estimation.

Given a sample $S = \{x_i\}_{i=1\dots N}$ from a distribution with density function $p(x)$, an estimate $\hat{p}(x)$ of the density at x can be calculated using

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} K_\sigma(x - x_i) \quad (1)$$

Where K_σ is a kernel function (sometimes called a "window" function) with a bandwidth (scale) σ such that: $K_\sigma(t) = \frac{1}{\sigma} K\left(\frac{t}{\sigma}\right)$. It is non-negative, integrates to one and has mean zero.

The Gaussian kernel function has been particularly used in this research as a function that weights included points. It was preferred because of its continuity, differentiability and locality properties *(2)*.

Data from one month at the initial stage, and for longer time periods at a later stage were analyzed. For each day kernel density values were estimated and used for the calculation of the average day of the month to which they relate, according to the equation:

$$\bar{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \hat{p}(x_i) \quad (2)$$

Where: N is the number of days that are included in the generalized average and $\hat{p}(x_i)$ is the kernel density values that correspond to one day. The kernel density values of a day are represented as a two dimensional array. The step of the analysis grid is defined according to the scale of the studied area. In our analysis, each cell of the kernel grid is also called "pixel".

Figure 3a shows the monthly average demand of taxi pick-ups in the area of Manhattan as it was produced from our pixel-by-pixel time-series model. Blue shades imply areas with high average demand while lighter shades (yellow to white) imply lower demand. The presented figure is a merge of the analysis results in Python and a background image exported from OpenStreetMaps. This generalized depiction already gives us enough information about the average dynamics of the studied area and it will help us evaluate on a more concrete basis the results obtained from individual day trips analysis.

Individual Days Scanning

Knowing the distribution of demand in the area of interest for an average day, we proceed to the next stage of individual days scanning. We are interested in finding days where demand representation differs significantly from the "normality" that the average day represents. The comparison phase takes into account only the kernel density values, and uses the Z-Score formula:

$$Z_{Diff,i} = \frac{\hat{p}(x_i) - \hat{p}(x)}{\sigma} \quad (3)$$

Where: i is the examined day and σ is the standard deviation of the average day. Through the Z-Score data transformation, each kernel density value is given in units of how many standard deviations it is from the mean value, and consequently how far from the average day demand levels. Maximum differences of each day will be used for the next stage of analysis which is their spatial localization and explanation. A depiction of demand differences is shown in Figure 3b, where red shades represent higher demand from the prior average analysis, while blue the opposite effect.

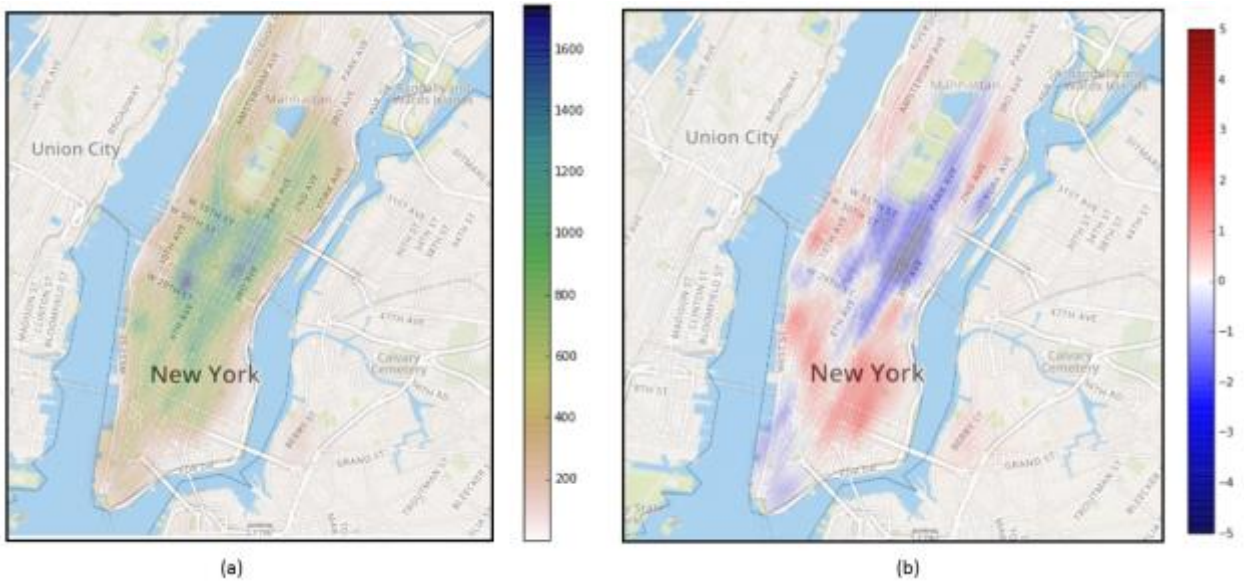


FIGURE 3: Kernel Density analysis depiction of (a) the average day (b) demand differences

Anomalies explanation

The Internet is a valuable resource for extracting information about special events, such as their location, duration and their popularity through Facebook likes or Google trends. Therefore, it could be a very useful tool for the level of assessment we are interested in this research.

From the scanning procedure we are able to create a diagram that shows the most significant outliers captured for each day of the time period we are interested in. There is no well-defined threshold above which we can identify a demand side anomaly. Therefore, in order to cope with all these observed demand fluctuations, days with Z-Score greater than 2 are initially selected for further analysis.

The information collected from the previous stage includes not only the days that present abnormal behavior, but also the location where the phenomenon of significantly high or low demand was noticed. The location was registered by identifying the "pixel" that shows the highest difference from the average picture. As a result, utilizing the above information we can find through Google search if a special event was held in this region. The generative procedure is presented in Figure 4.

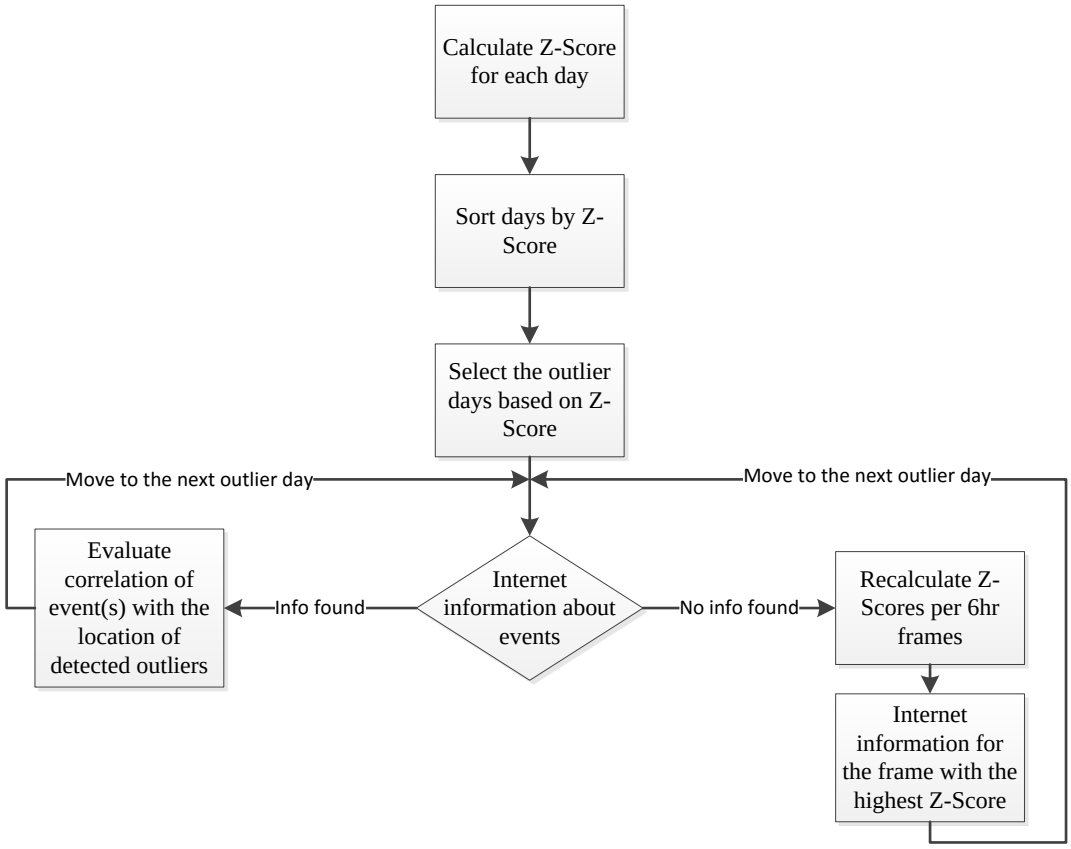


FIGURE 4 Methodology for anomalies detection and explanation

EXPERIMENTS

Data Description

In this research, Google BigQuery public datasets *(30)* were used as our main source of information. Particularly, the yellow taxi trips dataset in New York City since 2009 was explored. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. Data structure as well as its level of detail allows us to understand quite well the characteristics of taxi trips taken in a certain day. Through properly formatted SQL queries we were able to select our period of interest as well as the variables that we want to analyze in the context of our study.

Experimental Design

At the time of writing this document, Google BigQuery limits the maximum rate of incoming requests and enforces appropriate quotas on a per-project basis. There is also a maximum response size of 128 MB compressed; therefore we decided to implement several experiments with a baseline of three months interval average day. Each implemented query created a csv file, which was saved and later exported from the Google Cloud Storage.

Our research was focused on 2013 and 2015 monthly data, around the area of Manhattan, the most densely populated borough of New York City. Each month, numerous events, festivals, parades are organized. Several unexpected incidents further degrade the already low level of serviceability of city's road network, such as accidents, road works etc.. Road networks are also vulnerable to natural disasters such as floods and blizzards, which can adversely affect the travel

on the network that remains intact after an event. Our goal is to detect demand related anomalies, analyze and correlate them with scenarios like large sports games, concerts, religious festivities, demonstrations. Studying and analyzing vulnerability of road networks will help in prioritizing the planning and budgeting and also will be useful in preparing emergency response plans.

Results

From the implemented analysis on taxi trips datasets in 2013 and 2015, several whole-day events and small scale events were identified. The most characteristic examples are presented below.

Whole Day Events

High differences for a whole day were generally noticed when there is a public holiday, or the city faces extreme weather conditions *(31)*. By implementing the methodology described in detail in the previous section, we are able to identify high deviations from the average day whenever a blizzard affected the city, as well as on the Memorial Day.

Figure 5 shows the Z-Score values of each day in February 2013, compared with the average day of the corresponding month. February 9th presents the highest difference of the month. Using Google search it can be easily found that a northeast US blizzard affected the area of Manhattan on that day, therefore taxi trip counts and distribution is justifiably different from the average day.

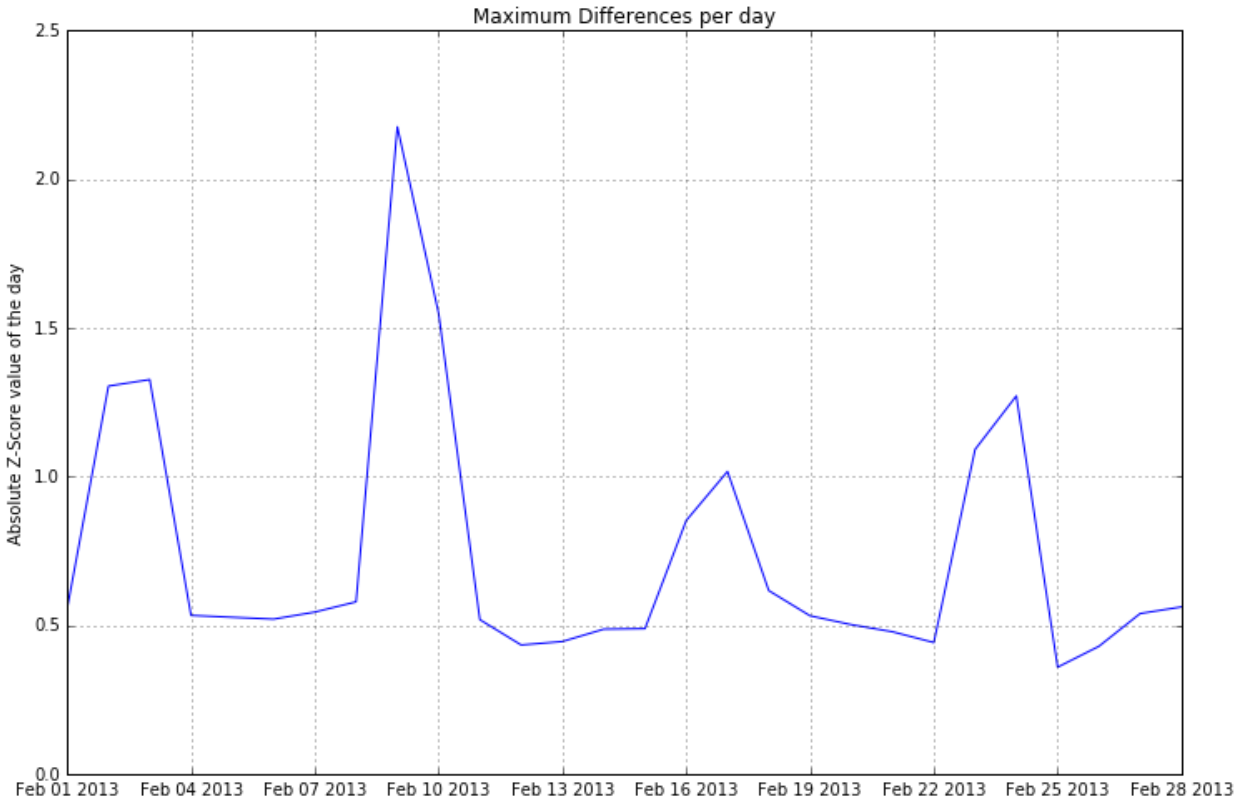


FIGURE 5 Absolute Z-Score values graph for February 2013

But since extreme weather conditions are clearly noticeable, and affect the entire region of a city, we further focused on events that mostly show unexpected changes and traffic anomalies

that transportation authorities and drivers could not easily explain and predict. Remaining in the category of all-day events, special interest presents the Memorial Day. Citizens and authorities know in advance that taxi trips will be lower compared to an average Monday, because it is a public holiday. From an initial analysis of the data this hypothesis could be easily proved (see Figure 6a). Taxi trip counts decline is evident for this public holiday.

By further examining the differences of pick up points distributions from the average day of May (Figure 6b), a significant increase in traffic is observed around the New York Penn Station (Figure 6c). This result can be justified by the fact that many people choose to travel for the Memorial Day weekend. Some of the top destinations based on American Automobile Association (AAA) travel agency sales and AAA.com are Orlando, Myrtle Beach, Washington D.C. and Miami. Each of these destinations is served by train trips that departure from Pennsylvania station, therefore it is common that many people will choose, after the end of their vacation, this station as a way of reaching the center of Manhattan and their destination by taxi afterwards.

Small Scale Events

Studying taxi pick up points for a whole day may not show significant changes in their distribution, because time intervals with higher fluctuations are combined with those that present low mobility patterns, such as night hours. Subsequently, our study focused thereafter mostly on morning and afternoon pick hours, where people choose to participate in events, such as a sport game or a concert.

By studying the three months period of September – November 2015, several traffic anomalies were noticed by selecting the Z-Score values that significantly deviated from the average value.

31 October 2015 – New York City Halloween Parade

On Saturday October 31st we noticed a significantly high Z-Score value compared to the average Saturday of the three months' period. By plotting the location of the highest value, and by further searching on the internet, we easily found that a Halloween parade was organized that day. Figure 1 shows how the distribution of taxi pick up points was developed after parade's end. The intense activity was transferred to the walkways and to the 7th Avenue, at the left side of the parade. The scale of the phenomenon around parade's route is consistent with the distance that someone may need to walk in order to find a taxi.

21 September 2015 – Metropolitan Opera Opening Night Gala

September was a month with several outliers from the average day of the three months' period. On September 21st a Z-Score value of 1.2 led to a further analysis for the time period 21:00-23:30. The results are shown in Figure 7a. The localization of the "pixel" that presented the highest Z-Score value (red star in Figure 7a) indicated that the metropolitan opera could be the explanation of this high demand. Eventually, it was proven, that on 21st September, the Metropolitan Opera's 'Othello' Opening Night Gala was organized, an event that attracts the spotlight and therefore the attendance of many prominent persons.

Figure 7a depicts also really low demand close to the Pennsylvania station. By further investigating the result, it was found that on 7th September the corresponding demand in this area was significantly high (Figure 7b), thus affecting the average corresponding day. Several people probably wanted to change mode after their arrival at this central station from numerous events organized in the greater area of NYC (Indian carnival, the beginning of NYC Broadway week and the US Open).

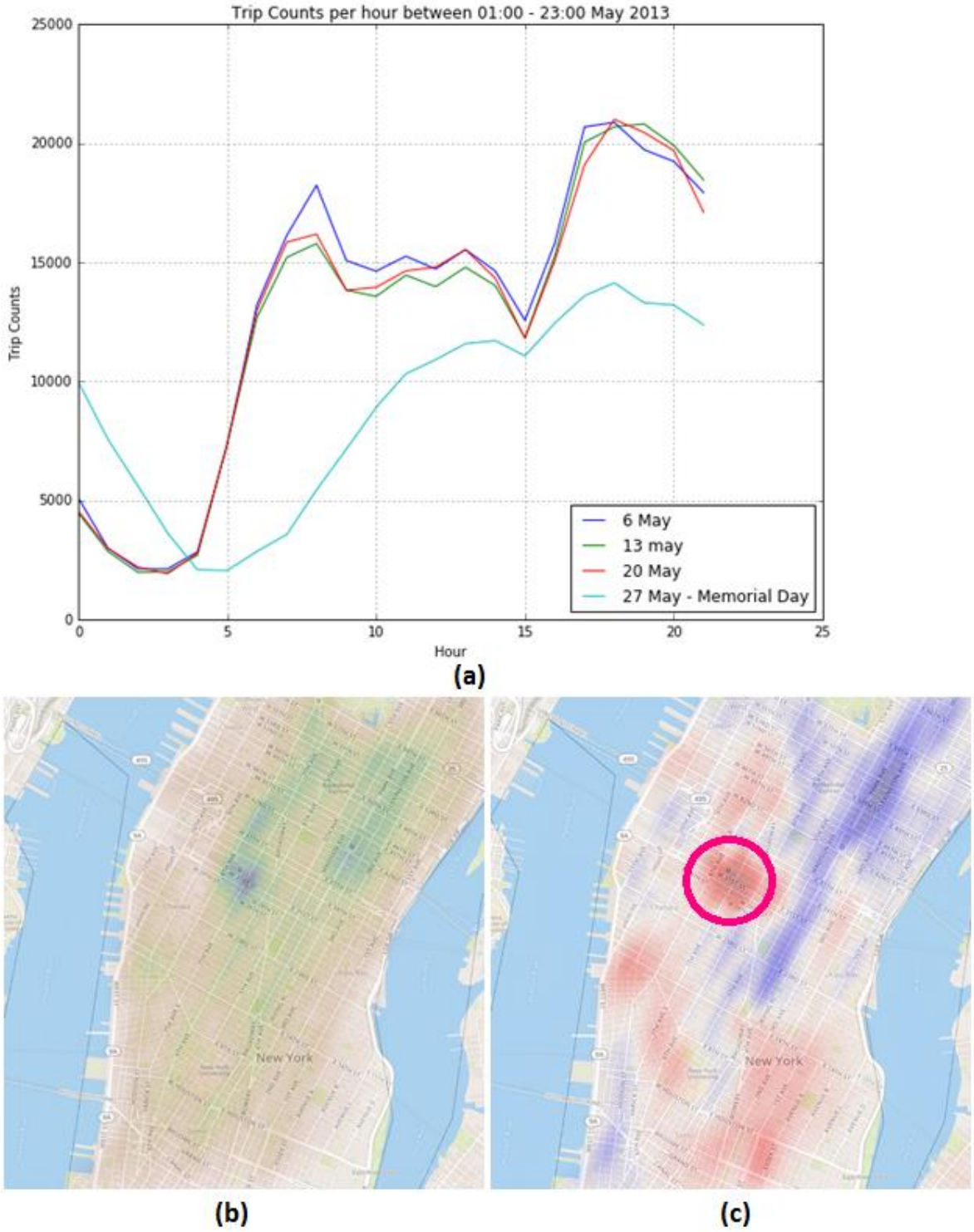
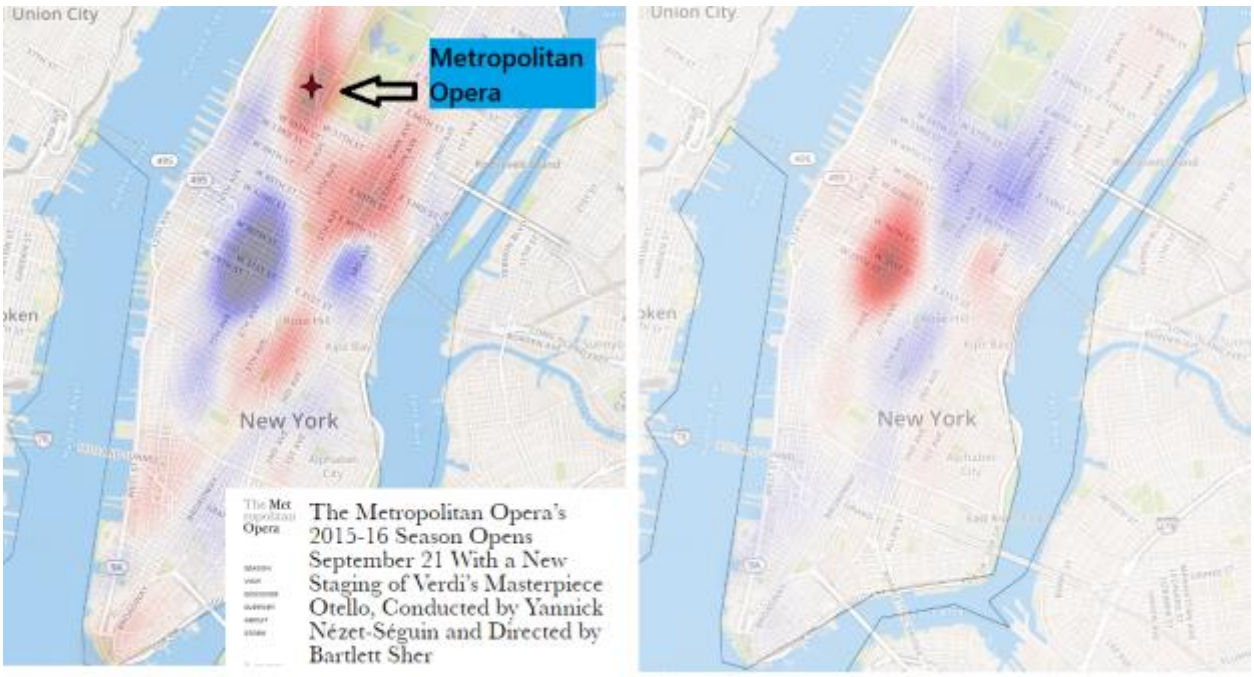


FIGURE 6: (a) Taxi Trip Counts per hour for each Monday of 2013 (b) Demand depiction on an average day in May, (c) demand on Memorial Day



(a) (b)

FIGURE 7. Depiction of demand differences on (a) 21st October, (b) 7th October

24 September 2015 – Pope Francis Visit

High differences on taxi pick up points distributions were noticed on 24th September between the time interval 15:00 – 21:00. The Z-Score value for that day was -2.155. An online search showed that Pope Francis visited New York City that day, and more specifically the St Patrick's Cathedral for an evening prayer. The negative sign of Z-Score indicates an overall reduction of taxi pick-ups in the hotspots of the average day; a conclusion which is justified by the scheduled road closures in many parts of the city for security reasons. Taxi pick up points are increased around the area of the cathedral and not closed to it, because as Figure 8 shows, all transits through the roads around it were forbidden.

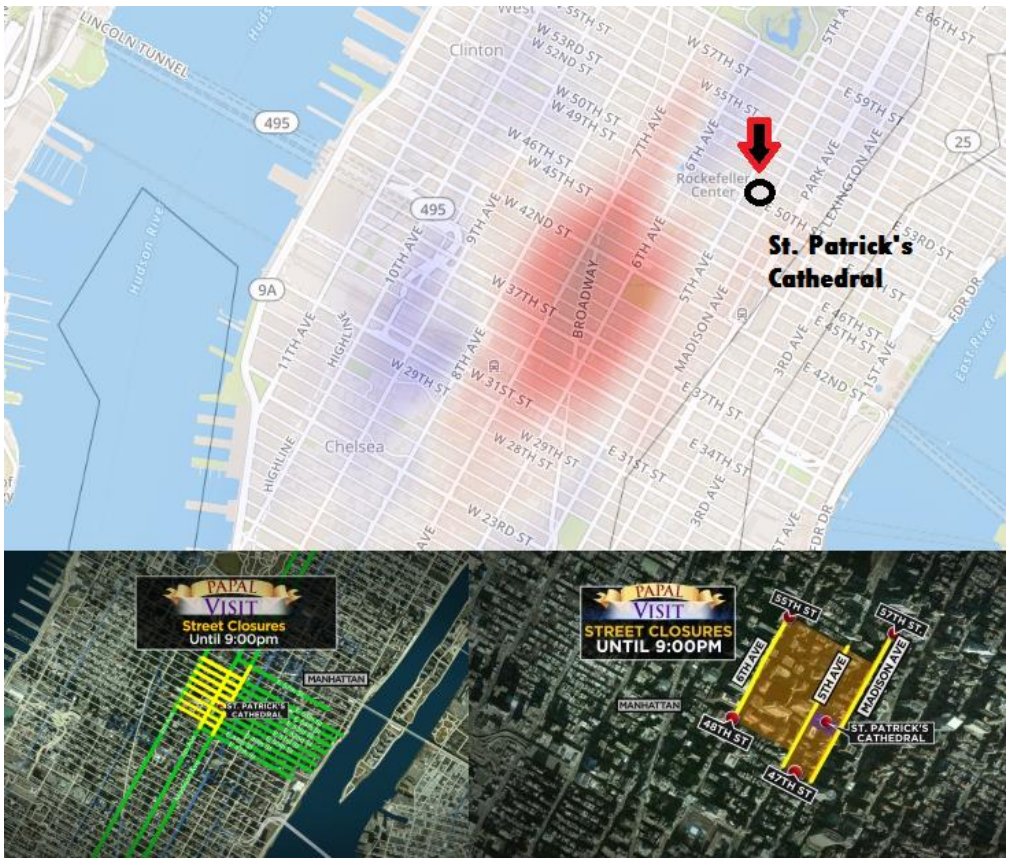


FIGURE 8 High Demand during Pope's presence at St. Patrick's Cathedral

CONCLUSION AND FUTURE WORK

Transportation networks present several anomalous situations of particular interest and merit. Through this research, we presented a methodology that identifies traffic anomalies on traffic networks and correlates them with special events using internet data. We evaluated our system with the Yellow taxi trips dataset in New York City during 2013 and 2015. We defined initially the "normality" baseline and thereunder we studied individual days' demand patterns for outliers' detection. Our approach enabled us to detect demand fluctuations, analyze and correlate them with disruptive events scenarios like extreme weather conditions, public holidays, religious festivities and parades. Using kernel density analysis, the affected areas as well as the significance of the observed differences compared to the average day are depicted. Based on the investigation results, it becomes distinguishable how a special event affects the spatial and temporal traffic flow in a studied region.

The present study is not yet about automatically searching from the web for a random event, instead it's about getting the first building blocks for this endeavor: automatically detect time, location, and magnitude of such events for real-time explanation of traffic congestions or road closures which are highly correlated with them. Utilizing our prior knowledge, we want to monitor and predict in real-time the state of the transportation network in non-recurrent scenarios, such as the events described in the research, and provide real-time information and guidance to travelers and transportation administrators. Future research includes the application of information retrieval techniques to automatically capture relevant documents that explain abnormal conditions of the transport network identified by anomaly detection algorithms, and the use of natural language processing and popularity estimation techniques to extract contextual features that can be

incorporated in transport prediction models, thereby making them context-aware and more adaptive to the demand.

REFERENCES

1. Pereira, F.C., Rodrigues, F. and Ben-Akiva, M. Using data from the web to predict public transport arrivals under special events scenarios. *Journal of Intelligent Transportation Systems*, 19(3), 2015, pp.273-288.
2. Pereira, F.C., Bazzan, A.L. and Ben-Akiva, M.E. The Role of Context in Transport Prediction. *IEEE Intelligent Systems*, 29(1), 2014, pp.76-80.
3. Perallos, A. Intelligent Transport Systems: Technologies and Applications. *John Wiley & Sons*., 2015
4. Parkany, E. and Xie, C. A complete review of incident detection algorithms & their deployment: what works and what doesn't (No. NETCR 37, NETC 00-7), 2015.
5. Li, X., Li, Z., Han, J. and Lee, J.G. Temporal outlier detection in vehicle traffic data. In *2009 IEEE 25th International Conference on Data Engineering*, IEEE, 2009, pp. 1319-1322.
6. Sheu, J.B. A sequential detection approach to real-time freeway incident detection and characterization. *European Journal of Operational Research*, 157(2), 2004, pp.471-485.
7. Chen, A., Khorashadi, B., Chuah, C.N., Ghosal, D. and Zhang, M. Smoothing vehicular traffic flow using vehicular-based ad hoc networking & computing grid (VGrid). In *2006 IEEE Intelligent Transportation Systems Conference*, IEEE, 2006, pp. 349-354.
8. Barria, J.A. and Thajchayapong, S. Detection and classification of traffic anomalies using microscopic traffic variables. *IEEE Transactions on Intelligent Transportation Systems*, 12(3), 2011, pp. 695-704.
9. Pan, B., Zheng, Y., Wilkie, D. and Shahabi, C. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2013, pp. 344-353.
10. Liu, W., Zheng, Y., Chawla, S., Yuan, J. and Xing, X. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 1010-1018.
11. Christoforou, Z., Corbille, E., Farhi, N. and Leurent, F. Managing planned disruptions of mass transit systems. In *TRB Transportation Research Board-95th annual meeting* Washington D. C., 2016. pp. 13.
12. Castro, P.S., Zhang, D. and Li, S. Urban traffic modelling and prediction using large scale taxi GPS traces. In *International Conference on Pervasive Computing*, Springer Berlin Heidelberg, 2012, pp. 57-72.
13. Lee, R. and Sumiya, K. Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, 2010, pp. 1-10.
14. Becker, H., Naaman, M. and Gravano, L. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 291-300.
15. Nichols, J., Mahmud, J. and Drews, C. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, 2012, pp. 189-198.
16. Pereira, F.C., Rodrigues, F., Polisciuc, E. and Ben-Akiva, M. Why so many people? Explaining Nonhabitual Transport Overcrowding With Internet Data. *IEEE Transactions on Intelligent Transportation Systems*, 16(3), 2015, pp.1370-1379.

17. Yardi, S. and Boyd, D. Tweeting from the Town Square: Measuring Geographic Local Networks. In *ICWSM*, 2010
18. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web*. 2010, pp. 851-860.
19. Becker, H., Iter, D., Naaman, M. and Gravano, L. Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012, pp. 533-542.
20. Watanabe, K., Ochi, M., Okabe, M. and Onai, R. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2541-2544.
21. Abdelhaq, H., Sengstock, C. and Gertz, M. Eventweet: Online localized event detection from twitter. In *Proceedings of the VLDB Endowment*, 6(12), 2013, pp.1326-1329.
22. Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X. and Hu, C. Crowdsourcing based description of urban emergency events using social media big data. In *IEEE Transactions on Cloud Computing*. 2016, pp. 1–1.
23. Zhang, J. Smarter outlier detection and deeper understanding of large-scale taxi trip records: a case study of NYC. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 2012, pp. 157-162.
24. Bu, Y., Chen, L., Fu, A.W.C. and Liu, D. Efficient anomaly monitoring over moving object trajectory streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 159-168.
25. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G. and Barabási, A.L. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22), 2008.
26. Chandola, V., Banerjee, A. and Kumar, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 2009, p.15.
27. Blei, D.M., Ng, A.Y. and Jordan, M.I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 2003, pp.993-1022.
28. Pereira, F.C., Rodrigues, F. and Ben-Akiva, M. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37, 2013, pp.177-192.
29. Quercia, D. and Saez, D. Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing*, 13(2), 2014, pp.30-36.
30. *BIGQUERY*. https://cloud.google.com/bigquery/. Accessed Aug. 1, 2016
31. Ferreira, N., Poco, J., Vo, H.T., Freire, J. and Silva, C.T. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2013, pp.2149-2158.